Smart Media Journal / Vol.15, No.1 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.1.42

제조업 AI 확산을 위한 On-Premise RAG의 실증적 타당성 분석 연구

(An Empirical Study on the Feasibility Analysis of On-Premise RAG for AI Diffusion in Manufacturing)

유재석*, 홍아름**
(Jaeseok You, Ahreum Hong)

요 약

제조 산업의 디지털 전환이 가속화됨에 따라 생산성 향상을 위한 거대언어모델(LLM)의 도입 필요성이 증대되고 있다. 그러나 중소·중견 제조기업은 자본과 인프라 제약으로 인해 클라우드 기반 상용 LLM의 데이터 보안 위험과 지속적인 고비용 구조를 주요 도입 장벽으로 인식하고 있다. 본 연구는 기술 - 조직 - 환경(TOE) 프레임워크 관점에서 이러한 한계를 극복하기 위한 대안으로 On-Premise 기반 검색 증강 생성(RAG) 시스템을 제안하고, 그 실증적 타당성을 검증하였다. 이를 위해 외부 지식 참조가 없는 제로샷 모델(NO RAG), 상용 클라우드 모델기반(GPT RAG), 그리고 오픈소스 모델기반 (OPEN MODEL RAG)를 구축하여 비교 실험을 수행하였다. 평가는 정량적 성능 지표(ROUGE, BLEU, BERTScore), 자원 효율성(토큰 사용량, 응답 시간), 그리고 LLM-as-Judge 기반 정성 평가(유용성, 간결성, 관련성)를 포함한 다차원적 기준으로 이루어졌으며, ANOVA와 Tukey의 HSD 사후 검정을 통해 통계적 유의성을 검증하였다. 분석 결과, 기술적 맥락에서 OPEN MODEL RAG는 GPT RAG와 통계적으로 대등한 정량적 성능을 보였으며, 정성적 평가에서는 유용성, 간결성, 관련성 측면에서 상용 모델을 유의미하게 상회하였다. 조직적 맥락에서는 응답 지연의 한계가 존재하였으나, 지속적인 토큰 사용 비용을 제거함으로써 운영 효율성을 확보하였다. 환경적 맥락 에서는 데이터의 외부 유출을 원천적으로 차단하여 제조기업의 보안 및 규제 요구사항을 충족함을 확인하였다. 본 연구는 보안이 중요한 제조 전문 도메인에서 오픈 소스 기반 On-Premise RAG가 비용 효율성과 신뢰성을 동시에 확보할 수 있는 기술경영학적 최적 대안이 될 수 있음을 실증적으로 제시한다.

■ 중심어 : On-Premise ; 제조업 AI ; 데이터 보안 ; TOE ; LLM-as-Judge

Abstract

This study empirically validates the feasibility of On-Premise Retrieval-Augmented Generation (RAG) as a techno-managerial alternative for SMEs facing security and cost barriers in AI adoption. Applying the TOE (Technology-Organization-Environment) framework, we conducted a comparative analysis between a Base Model (Zero-shot), GPT RAG, and an open-source-based On-Premise RAG. Results indicated that the On-Premise RAG not only matches commercial models in quantitative performance but also superiorly fulfills qualitative criteria such as usefulness and relevance by utilizing a systematic knowledge-base approach. While exhibiting inherent latency, the system optimizes organizational efficiency by eliminating recurring token costs and ensures environmental security by fundamentally preventing external data leakage. Ultimately, On-Premise RAG provides a reliable, cost-effective solution for maintaining data sovereignty in specialized manufacturing domains.

■ Keywords : On-Premise ; Manufacturing AI ; Data Security ; TOE ; LLM-as-Judge

Ⅰ. 서 론

대규모 언어 모델(Large Language Models, LLM)을 중심으로 한 생성형 인공지능 기술은 전 산업의

* 정회원, 경희대학교 테크노경영대학원 AI 기술경영학과, 텔스타 연구소장
** 정회원, 경희대학교 테크노경영대학원 AI 기술경영학과, 교신저자
이 논문은 2025년도 경희대학교 테크노경영대학원 AI 기술경영학과 석사 논문 연구임.

접수일자 : 2025년 12월 29일
수정일자 : 2026년 01월 23일
게재확정일 : 2026년 01월 26일
교신저자 : 홍아름 e-mail : arhong@khu.ac.kr

업무 방식과 의사결정 구조를 빠르게 변화시키고 있다. 문서 분석, 지식 검색, 공정 매뉴얼 자동화 등 다양한 영역에서 LLM의 활용 가능성이 제시되며, 제조 산업 역시 생산성 향상과 운영 효율성 제고를 위한 핵심 기술로 주목받고 있다. 그러나 이러한 기대와 달리, 제조 산업 특히 중소·중견 기업을 중심으로 LLM 기반 인공지능의 실질적인 도입은 제한적으로 이루어지고 있다. 기존 연구와 산업 보고서에 따르면, 제조업에서의 AI 활용은 소수의 선도 기업에 집중되어 있으며, 다수의 기업은 여전히 도입 초기 단계에 머물러 있는 것으로 나타났다[1].

본 연구는 이러한 도입 지연 현상을 기술 - 조직 - 환경(Technology - Organization - Environment, TOE) 프레임워크 관점에서 분석하고, 제조 기업이 직면한 구조적 장벽을 체계적으로 규명하고자 한다. 첫째, 기술적 맥락(Technology Context)에서 LLM은 학습 데이터 시점의 한계로 인한 지식 단절과 사실과 다른 정보를 생성하는 환각(hallucination) 문제를 내포하고 있어, 높은 정확성과 신뢰성이 요구되는 제조 현장에 적용하는 데 제약이 존재한다. 둘째, 조직적 맥락(Organization Context)에서 상용 클라우드 기반 LLM 활용은 핵심 기술 및 데이터 유출에 대한 보안 우려와 함께, 사용량에 비례하여 지속적으로 증가하는 토큰 비용 부담을 초래함으로써 조직 차원의 수용성을 저하시킨다. 셋째, 환경적 맥락(Environment Context)에서는 데이터 주권 확보와 소버린 AI에 대한 요구가 확산됨에 따라, 특정 글로벌 빅테크 기업에 대한 기술 종속을 최소화하고 독립적인 AI 운영 환경을 구축하려는 요구가 증대되고 있다.

이러한 문제의식 하에 본 연구는 외부 네트워크와 분리된 On-Premise 환경에서 운영 가능한 오픈소스 모델 기반 검색 증강 생성(Retrieval-Augmented Generation, RAG) 시스템에 주목한다. 오픈소스 RAG는 기업 내부 데이터만을 활용하여 응답을 생성함으로써 정보의 외부 유출을 원천적으로 차단할 수 있으며, 사용량 기반 과금 구조에서 벗어나 비용 예측 가능성과 운영 효율성을 동시에 확보할 수 있다. 또한 외부 지식 검색을 결합함으로써 LLM의 환각 문제를 완화하고, 제조 도메인에 특화된 신뢰도 높은 응답을 생성할 수 있는 잠재적 대안으로 평가된다.

따라서 본 연구의 목적은 오픈소스 모델을 활용한 On-Premise RAG 시스템이 고성능 상용 LLM 기반 시스템의 실질적인 대안이 될 수 있는지를 실증적으로 검증하는 데 있다. 이를 위해 본 연구는 RAG를 적용하지 않은 제로샷 모델(NO RAG), 상용 클라우드 LLM을 적용한 RAG(GPT RAG), 오픈소스 LLM을 적용한 On-Premise RAG (OPEN MODEL RAG)의 세 가지 아키텍처를 구축하고, 지식 베이스 기반으로 구성된 질의응답 데이터셋을 동일하게 적용하여 성능을 비교·분석하였다. 평가는 정량적 품질 지표(ROUGE-L, BLEU, BERTScore), LLM-as-Judge를 활용한 정성적 평가, 그리고 자원 효율성(토큰 사용량, 응답 속도)을 포함한 다차원적 기준을 통해 수행되었다. 이러한 실증 분석을 통해 중소 제조 기업을 위한 현실적인 LLM 도입 전략을 제시하고자 한다.

Ⅱ. 본 론

1. 이론적 배경과 선행 연구

가. 기술 도입 의사결정 프레임워크 TOE

TOE 프레임워크는 Tornatzky & Fleischer(1990)[2]가 제안한 이론으로, 기업이 새로운 기술 혁신을 채택하고 구현하는 과정에 영향을 미치는 요인을 다각도로 분석하는 데 널리 활용된다. TOE 프레임워크는 기술 도입을 단순히 기술적 우수성으로만 판단하지 않고, 조직적 맥락과 환경적 맥락을 통합적으로 고려한다는 점에서 제조 기업의 AI 도입 장벽을 분석하는 데 적합한 틀을 제공한다.

기술적 맥락은 기업이 도입하려는 기술의 특성과 가용성을 의미한다. 생성형 AI 도입에 있어 이는 모델의

성능, 정확성, 그리고 기존 시스템과의 호환성을 포함한다. 조직적 맥락은 기업의 자원, 규모, 비용 수용 능력 등을 포괄하며, 특히 중소 제조 기업의 경우 고비용의 인프라 구축이나 지속적인 운영 비용이 주요한 의사결정 변수로 작용한다. 환경적 맥락은 산업 내 경쟁, 정부 규제, 데이터 보안 요구사항 등을 의미한다. 최근 데이터 주권과 보안 규제가 강화됨에 따라, 제조 기업은 기술 도입 시 외부 플랫폼 종속성이나 데이터 유출 위험을 심각하게 고려해야 하는 상황에 직면해 있다. 본 연구는 이러한 세 가지 맥락을 기반으로 On-Premise RAG 시스템의 도입 타당성을 분석한다.

나. LLM의 한계와 RAG의 등장

대규모 언어 모델(LLM)은 방대한 데이터를 학습하여 뛰어난 언어 이해 및 생성 능력을 보여주지만, 제조 도메인과 같이 높은 신뢰성이 요구되는 현장에 바로 적용하기에는 몇 가지 치명적인 한계를 가진다.

첫째, 환각 현상이다. LLM은 확률적 모델로서 사실이 아닌 정보를 그럴듯하게 생성하는 경향이 있어, 정확한 수치와 절차가 중요한 제조 공정 가이드에서 심각한 오류를 유발할 수 있다[3]. 둘째, 지식 단절 및 내부 데이터 접근 불가 문제이다[4]. 범용 LLM은 학습 시점 이후의 최신 정보를 알지 못하며, 기업 내부의 비공개 데이터(설비 매뉴얼, 공정 노하우 등)는 학습 데이터에 포함되지 않아 구체적인 현장 질의에 답변하는 데 한계가 있다[4].

이러한 문제를 해결하기 위해 RAG 기술이 등장하였다. RAG는 LLM이 답변을 생성하기 전에 신뢰할 수 있는 외부 지식 베이스에서 관련 정보를 먼저 검색하고, 이를 프롬프트에 포함하여 생성을 수행하는 방식이다. 이는 모델을 재학습하지 않고도 최신 정보와 내부 데이터를 반영할 수 있어, 비용 효율적으로 LLM의 정확도를 높이는 핵심 기술로 주목받고 있다[5].

다. On-Premise RAG의 필요성

첫째, 기술적 맥락에서 On-Premise RAG 시스템은 기존 LLM의 고질적인 한계인 환각 현상을 효과적으로 완화한다. 외부 지식 베이스를 활용하여 신뢰할 수 있는 정보를 참조함으로써 정확한 수치와 절차가 중요한 제조 현장의 특수성을 반영하며, 제조 도메인에 특화된 고신뢰 응답을 생성할 수 있는 기술적 대안을 제공한다 . 이는 모델을 재학습하지 않고도 기업 내부의 설비 매뉴얼이나 공정 노하우 등 최신 데이터를 반영할 수 있어 기술적 유연성을 확보해 준다.

둘째, 조직적 맥락에서는 자원 할당과 장기적인 운영 비용의 예측 가능성이 핵심적인 의사결정 요인으로 작용한다. 상용 클라우드 기반 API 모델은 사용량에 비례하여 지속적으로 증가하는 토큰 기반 과금 체계를 가지고 있어 조직의 수용성을 저하시키고 운영 부담을 가중시킨다. 반면, 오픈소스 모델을 활용한 On-Premise 시스템은 초기 인프라 구축 비용 외에 추가적인 토큰 비용이 발생하지 않아 장기적인 운영 관점에서 경제적 타당성과 조직적 효율성을 동시에 달성할 수 있도록 돕는다.

셋째, 환경적 맥락에서는 데이터 보안 규제 준수와 기술 종속성 탈피를 통한 데이터 주권 확보가 강조된다. 제조 기업의 핵심 자산인 도면, 배합비, 공정 데이터가 외부 클라우드 서버로 전송되는 것은 산업 스파이 행위나 기밀 유출 위험 등 심각한 보안 리스크를 수반하며, 이는 기술 도입의 가장 큰 환경적 장벽이 된다 . On-Premise RAG는 모든 데이터 처리를 내부 폐쇄망에서 수행함으로써 보안 요구사항을 충족하며, 특정 빅테크 기업의 정책 변경이나 서비스 중단에 영향을 받지 않는 독립적인 기술 운영 환경을 구축할 수 있게 한다.

라. RAG 성능을 좌우하는 기술 요소(제조 문서 관점)

제조 기술 문서는 일반적인 텍스트와 달리 표, 수치, 도면 설명 등이 혼재된 복합적인 구조를 가진다. 따라서 단순한 RAG 구현이 아닌, 도메인 특성을 반영한 기술적 최적화가 필수적이다.

가장 중요한 요소는 청킹(Chunking) 전략이다. 문맥을 고려하지 않고 단순히 글자 수로 자르는 방식은 작업 절차나 수치 정보의 연결성을 끊어 검색 정확도를

떨어뜨린다[4]. 따라서 문서의 구조(문단, 소제목 등)를 인식하여 의미 단위로 분할하는 구조적 청킹이 요구된다. 또한, 검색 알고리즘의 경우, 제조 현장의 전문 용어(부품 번호, 약어 등)를 정확히 매칭하는 키워드 기반 검색(Sparse)과 문맥적 의미를 파악하는 벡터 검색(Dense)을 결합한 하이브리드 검색(Hybrid Search) 방식이 높은 성능을 발휘한다[7].

마. 평가 지표와 도입 의사결정의 연결

RAG 시스템의 도입 타당성을 검증하기 위해서는 단순한 정확도 측정을 넘어, 실무적 관점의 다각적인 평가가 필요하다. 본 연구에서는 이를 정량적 지표와 정성적 지표, 그리고 자원 효율성 지표로 구분하여 도입 의사결정의 근거를 마련한다.

정량적 지표로는 생성된 답변과 정답 문서 간의 텍스트 유사도를 측정하는 ROUGE[8], BLEU[9] 점수와 의미적 유사도를 평가하는 BERTScore[10]를 활용한다. 이는 시스템의 기술적 완성도를 판단하는 객관적 기준이 된다. 정성적 지표는 LLM-as-Judge[6] 기법을 활용하여 답변의 유용성, 관련성, 간결성을 평가한다. 이는 현장 작업자가 체감하는 실질적인 업무 지원 능력을 대변한다. 마지막으로 자원 효율성 지표인 응답 속도와 토큰 사용량은 조직적 비용과 직결되는 요소로, 기업이 고성능 상용 모델과 비용 효율적인 On-Premise 모델 사이에서 최적의 선택을 할 수 있도록 돕는 지표가 된다.

2. 연구 방법

가. 연구설계 개요

본 연구는 보안과 비용 효율성이 중시되는 제조 도메인 환경에서 On-Premise 기반 오픈소스 모델 RAG 시스템의 실질적인 활용 타당성을 검증하는 것을 목표로 한다. 이를 위해 실제 자동차 부품 조립 라인의 기술 정의서를 기반으로 지식 베이스를 구축하고, 응답 품질과 자원 효율성을 다차원적으로 분석하였다.

나. 데이터셋 구성 및 질문 설계

RAG 시스템의 지식 베이스는 차동 기어 조립 자동화 라인 기술 정의서를 활용하였다. 해당 문서는 총 17페이지(약 2.4 MB) 분량으로, 차동 기어의 투입(F10)부터 완제품 반출(F150)까지의 15개 핵심 공정을 다루고 있다. 문서의 구조는 공정번호, 설비 사양, 기계동작 순서, 품질 관리 데이터 등 9개의 표준 카테고리로 계층화되어 있으며, 표 및 구조화된 수치 데이터가 약 65%, 공정 설명 등 서술형 텍스트가 약 35%의 비율로 구성되어 있다. 특히 로봇의 반복 정밀도, 비전 센서의 분해능과 같은 고정밀 수치 정보와 공정 간 선후 관계를 정의하는 복합적인 기술 정보가 포함되어 있어, 단순 검색을 넘어선 정밀한 문맥 이해와 데이터 추출 성능을 평가하기에 적합한 복잡도를 갖추고 있다.

시스템의 다각적 성능 검증을 위해 설계된 100개의 평가용 질문셋은 제조 현장의 실질적인 정보 수요를 반영하여 다음과 같이 4가지 유형으로 배분하였다.

첫째, 단순 질의(40%)는 설비 모델명, 담당자, 수치 스펙 등 특정 단일 정보를 정확히 추출하는 검색 기본 성능을 평가한다. 둘째, 다중 컨텍스트 질의(30%)는 공정별로 분산된 설비 사양이나 부품 정보를 통합하여 답변하는 능력을 검증하며, 이는 RAG의 광범위한 문맥 참조 능력을 평가하는 지표가 된다. 셋째, 조건부 질의(20%)는 'NG 판정 시 배출 경로'나 '압입력 변화에 따른 품질 해석' 등 제조 공정 내의 의사결정 로직 이해도를 측정한다. 마지막으로, 추론 질의(10%)는 최종 검사 결과와 이전 조립 공정 간의 인과 관계를 파악하는 고차원적 추론 능력을 평가하는 데 중점을 두었다.

이러한 분포는 On-Premise RAG가 단순한 문서 검색기를 넘어, 제조 현장의 복잡한 의사결정을 지원하는 지능형 기술 지원 시스템으로서의 실효성이 있는지를 다각도로 검증하기 위함이다.

다. 시스템 구현

본 연구에서 제안하는 On-Premise RAG 시스템의 실증적 타당성을 검증하기 위해, 통제된 실험 환경 하

에 다음과 같이 시스템 아키텍처와 검색 파이프라인을 구축하였다.

(1) 시스템 구현 환경 및 하드웨어 사양

상용 LLM 기반의 제로샷 모델과 RAG 시스템 그리고 오픈소스 모델기반의 RAG 성능을 공정하게 비교하기 위해 표준화된 소프트웨어 스택을 구축하였다. On-Premise 시스템인 OPEN MODEL RAG는 EEVE-Korean-Instruction-10.8B 모델을 기반으로 하며, 다음과 같은 하드웨어 환경에서 운영되었다.

<표 2.1> 하드웨어 사양

하드웨어	OS	Ubuntu 24.04.1 LTS
	CPU	Intel Core i9-10900X (20 threads @ 4.60GHz)
	GPU	2 x NVIDIA GeForce RTX 2080 Ti
	RAM	128 GiB

(2) 의미 구조 보존을 위한 데이터 전처리 및 인덱싱 전략

제조 기술 문서는 표, 수치, 공정 순서가 밀집된 복합적인 구조를 가지므로, 단순한 텍스트 분할은 정보의 단절을 초래할 수 있다. 이를 방지하고 의미 구조를 보존하는 인덱싱을 구현하기 위해 다음과 같은 청킹 전략을 적용하였다. RecursiveCharacterTextSplitter를 활용하여 문맥의 연속성을 확보하고, 실제 구현 코드에 기반하여 청크 크기는 800자, 청크 중첩은 100자로 설정하였다. 의미 단위 분할을 위한 구분자를 ["\n\n", "\n", " ", ""] 순으로 적용하여, 이중 줄바꿈(\n\n)을 통한 공정 간 경계와 단일 줄바꿈(\n)을 통한 표 내 항목 정보를 최우선적으로 보존하도록 설계하였다.

(3) 하이브리드 검색 상세 사양 및 최적화

제조 현장의 전문 용어(부품 번호, 약어 등)에 대한 정확한 매칭과 문맥적 의미 파악을 동시에 달성하기 위해 하이브리드 검색 방식을 채택하였다. 키워드 기반 검색인 BM25와 벡터 유사도 기반 검색인 FAISS를 EnsembleRetriever로 결합하였다. BM25와 FAISS의 가중치 비율을 0.5:0.5로 동일하게 설정하여 고유 명사 매칭과 문맥 유사성 검색의 균형을 도모하였다. 최종 LLM에 전달되는 검색 문서 개수(Top-K)는 3개로 제한하였다. 이는 On-Premise 환경에서의 제한된 GPU 자원을 고려하여 응답 지연을 최소화하고 정보 밀도를 높이기 위한 설정이다.

본 연구의 실험 설계 단계에서 임베딩 모델로 OpenAI의 text-embedding-3-small을 채택했다. 상용 모델과 오픈소스 모델간의 생성 성능을 동일한 지식 품질 기준 하에 객관적으로 비교하기 위한 실험적 통제의 결과이다. 비록 인덱싱 과정에서 외부 API를 사용한 점이 본 연구가 지향하는 보안성과 데이터 주권 관점에서 일부 상충하는 한계가 있음을 인지하고 있으나, 이는 검색된 컨텍스트를 해석하고 응답을 생성하는 LLM 본연의 추론 능력을 공정하게 평가하기 위함이었다. 실제 제조 현장의 보안 요구사항을 완벽히 충족하기 위해서는 본 연구에서 검증된 아키텍처를 기반으로 하되, 임베딩 모델을 허깅페이스 기반의 로컬 배포형 모델(예: BGE-M3,)로 대체함으로써 데이터의 외부 유출을 원천 차단하는 최종 아키텍처 완성이 가능하다.

<표 2.2> 실험모델 구성 비교

구분	NO RAG	GPT RAG	OPEN MODEL RAG
LLM	GPT-4o-mini	GPT-4o-mini	EEVE-Korean-10.8B
파라메타	비공개	비공개	108억
지식	Zero-shot	참조	참조

라. 평가 절차 및 통계 분석

RAG의 실효성 검증을 위해 비교 실험군을 다음과 같이 정의하였다. 첫째, 제로샷 모델(NO RAG)로 외부 지식 없이 GPT-4o-mini 자체 지식만 활용하여 도메인 지식 단절과 환각 현상을 측정한다. 둘째, 상용 RAG(GPT RAG)는 GPT-4o-mini에 본 RAG 파이프라인을 결합한 시스템이다. 셋째, 오픈소스 모델 RAG(OPEN MODEL RAG)는 On-Premise 기반 EEVE-10.8B 모델에 동일 파이프라인을 적용한 것이다. 모든 실험은 동일한 100개 질의셋을 통해 수행되었다. 생성된 응답은 (1) 정량 품질, (2) 정성 평가, (3) 자원 효율성의 3개 축에서 평가하였다. 정량 평가는 ROUGE-L, BLEU, BERTScore를 산출하여 시스템 간 응답 품질을 비교하였다.

정성 평가를 위해 LLM-as-Judge 기법을 도입하였으며, 평가 모델로는 뛰어난 추론 성능을 갖춘 GPT-4o를 활용하였다. 평가의 전문성과 일관성을 확보하기 위해 평가 모델에 제조 라인 및 품질 관리 전문가의 페르소나를 부여하였으며, 개별 평가 시에는 [질문], [정답], [시스템 답변]을 하나의 프롬프트 셋으로 구성하여 입력하였다. 이를 통해 모델이 전문가의 관점에서 정답 문서를 기준으로 시스템 응답의 정확성과 품질을 10점 척도로 정교하게 비교 판정하도록 설계하였다. 평가 기준의 분명함을 위해 세부 지표를 다음과 같이 정의하였다. 첫째, 유용성은 답변이 질문의 의도를 정확히 파악하여 작업자에게 실질적이고 구체적인 정보를 제공하는지를 측정한다. 둘째, 간결성은 핵심 정보를 누락하지 않으면서도 중복이나 장황한 설명을 배제한 정보 전달의 효율성을 평가한다. 셋째, 관련성은 답변이 검색된 기술 문서의 맥락과 논리적으로 일치하며 도메인 지식을 정확히 반영하고 있는지를 검증한다. 본 실험에서 GPT-4o를 평가자로 채택한 것은 모델의 성능을 신뢰도 높은 상용 지표와 대조하여 객관적으로 검증하기 위해서이다. 이는 연구 단계의 일회성 평가 과정일 뿐, 모든 데이터 처리가 내부 폐쇄망 내에서 이루어지는 실제 시스템 아키텍처의 보안 결함과는 무관하다.

자원 효율성 평가는 각 시스템의 토큰 사용량과 응답 시간을 기록하여 비교하였다. 토큰 사용량 집계는 한국어 전용 형태소 분석기인 Kiwi를 활용한 단일 측정 기준을 적용하였다. 시스템의 전체 처리 부하를 측정하기 위해 질문, 검색된 컨텍스트, 모델의 응답 텍스트 전체를 형태소 단위로 토큰화하여 합산하였다.

통계 분석은 지표별로 3개 시스템 간 평균 차이를 검정하기 위해 일원분산분석(one-way ANOVA)을 적용하였다. 분산분석에서 유의한 차이가 확인된 경우, 집단 간 차이를 구체적으로 확인하기 위해 Tukey HSD 사후검정을 수행하였다.

그림 1. 연구 흐름도

3. 연구 내용

가. 정량 품질(ROUGE-L, BLEU, BERTScore)

구축된 세 가지 시스템(NO RAG, GPT RAG, OPEN MODEL RAG)의 응답 품질을 정량적 지표인 ROUGE-L, BLEU, BERTScore를 통해 측정한 결과는 <표 3.1>과 같다.

<표 3.1> 품질 지표별 ANOVA 테스트 결과 (M±SD)

지표	NO RAG	GPT	OPEN	df	F	p
ROUGE-L	0.104 ±0.080	0.380 ±0.176	0.375 ±0.182	(2,297)	106.59	p<.001
BLEU	0.027 ±0.042	0.181 ±0.160	0.167 ±0.141	(2,297)	46.27	p<.001
BERTScore	0.639 ±0.058	0.763 ±0.060	0.756 ±0.064	(2,297)	132.41	p<.001

**유의수준은 α = .05로 판단하였음.*

분석 결과, ROUGE-L, BLEU, BERTScore 모든 지표에서 세 시스템(NO RAG, GPT RAG, OPEN MODEL RAG) 간 평균 차이가 통계적으로 유의하였다(ROUGE-L: F=106.59, p=1.274e−36; BLEU: F=46.27, p=3.202e−18; BERTScore: F=132.41, p=7.736e−42). 평균값을 보면 두 RAG 시스템(GPT RAG, OPEN MODEL RAG)은 NO RAG 대비 모든 지표에서 높은 값을 보였다.

사후검정 결과, GPT RAG와 OPEN MODEL RAG 간에는 ROUGE-L(p=0.964), BLEU(p=0.693), BERTScore(p=0.724) 모두 유의한 차이가 없었으나, 두 RAG 시스템은 NO RAG 대비 유의미하게 우수하였다.

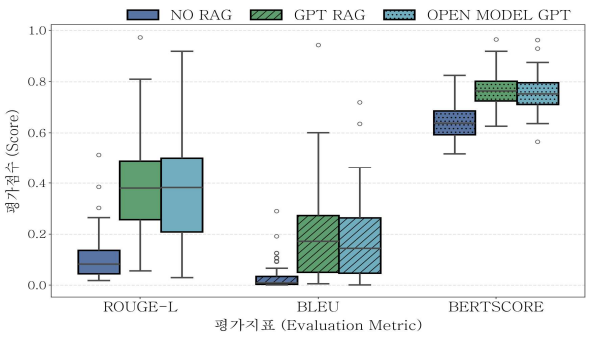

그림 2 모델별 정량 지표 평가 점수 분포

이러한 결과는 RAG 적용이 제조 기술 문서 기반 질의응답의 정량 품질을 유의미하게 개선한다는 점을 보여주며, 상용 기반과 오픈소스 모델 RAG의 정량 성능은 통계적으로 대등함을 시사한다.

나. 자원 효율성(토큰 사용량, 응답 시간)

구축된 세 가지 시스템(NO RAG, GPT RAG, OPEN MODEL RAG)의 응답 품질을 효율성 지표인 토큰 사용량, 응답시간을 통해 측정한 결과는 <표 3.2>과 같다.

<표 3.2> 품질 지표별 ANOVA 테스트 결과 (M±SD)

지표	NO RAG	GPT	OPEN	df	F	p
토큰사용량	222.24 ±104.89	149.650 ±33.37	86.510 ±52.79	(2,297)	92.89	p<.001
응답시간	5.680 ±3.15	5.220 ±1.32	30.720 ±16.54	(2,297)	223.80	p<.001

*유의수준은 α = .05로 판단하였음.

분석 결과, 토큰 사용량과 응답 시간 모두에서 세 시스템 간 차이가 통계적으로 유의하였다 (토큰: F=92.89, p=4.644e−32; 응답 시간: F=223.80, p=5.273e−60) 평균 토큰 사용량은 NO RAG(222.24) > GPT RAG(149.650) > OPEN MODEL RAG(86.510) 순으로 감소하였다. 사후검정에서는 OPEN MODEL RAG가 GPT RAG 대비 토큰 사용량이 유의미하게 적었고, GPT RAG도 NO RAG 대비 유의미하게 적었다.

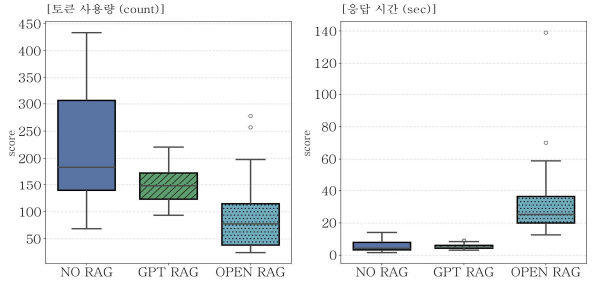

그림 3 모델별 효율성 지표 평가 점수 분포

반면 응답 시간은 OPEN MODEL RAG가 GPT RAG보다 유의미하게 길었으며, GPT RAG와 NO RAG 간에는 유의미한 차이가 없었다. 오픈소스 모델 RAG는 토큰 측면에서 효율적이나, 응답 속도 측면에서는 불리한 상충관계가 확인되었다.

다. 정성 평가(LLM-as-Judge: 유용성, 간결성, 관련성)

구축된 세 가지 시스템(NO RAG, GPT RAG, OPEN MODEL RAG)의 응답 품질을 정성적 지표인 유용성, 간결성, 관련성을 통해 측정한 결과는 <표 3.3>과 같다.

<표 3.3> 품질 지표별 ANOVA 테스트 결과 (M±SD)

지표	NO RAG	GPT	OPEN	df	F	p
유용성	4.240 ±2.23	6.290 ±3.36	7.850 ±2.52	(2,297)	43.47	p<.001
간결성	3.180 ±1.31	6.120 ±3.24	6.960 ±2.45	(2,297)	65.03	p<.001
관련성	4.190 ±2.04	6.700 ±3.24	7.970 ±2.46	(2,297)	53.73	p<.001

*유의수준은 α = .05로 판단하였음.

분석 결과, 유용성(F=43.471, p=2.759e−17), 간결성(F=65.033, p=3.764e−24), 관련성(F=53.731, p=1.210e−20)에서 세 시스템 간 차이가 모두 유의하였다. 평균값은 NO RAG 대비 GPT RAG와 OPEN MODEL RAG에서 전반적으로 높은 점수가 관찰되었으며, RAG 적용의 효과가 정성 지표에서도 확인되었다.

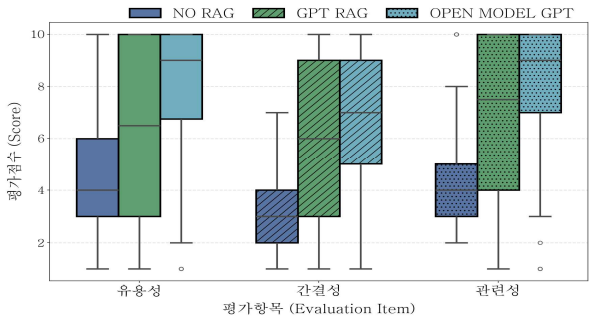

그림4 모델별 정성 지표 평가 점수 분포

사후검정 결과, 두 RAG 시스템은 NO RAG보다 모든 지표에서 유의미하게 높았고(p<.001), OPEN MODEL RAG는 GPT RAG보다 유용성(p=.0002),

간결성(p=.043), 관련성(p=.002)에서 유의미하게 높은 점수를 보였다. 정량 평가에서 상용과 오픈소스 모델 RAG가 대등했던 것과 달리, 정성 평가에서는 오픈소스 모델 RAG가 사용자 관점의 체감 품질에서 우위를 가질 수 있음을 시사한다.

본 연구의 결과는 RAG 환경에서 LLM의 역할이 방대한 파라미터에 의존하는 지식 저장고에서 주어진 지식 베이스를 논리적으로 재구성하는 효율적 추론기로 전이되고 있음을 시사한다.

이러한 성능 역전 혹은 대등함이 나타난 핵심 원인은 지식베이스에 대한 체계적 접근 방식에 있다. 본 연구에서 적용한 의미 구조 보존 인덱싱과 하이브리드 검색 기술은 LLM이 답변을 생성하기 위해 필요한 토픽 기반의 명확한 근거를 고순도로 정제하여 전달하였다. 이처럼 검색 단계에서 지식의 정교한 필터링이 선행될 경우, 오픈소스 모델은 상용 모델이 가진 방대한 범용 지식의 간섭 없이 제공된 컨텍스트에만 집중하여 추론을 수행하게 된다.

특히 정성 평가에서 오픈소스 모델이 우위를 보인 점은, 모델에 명확한 근거가 전달되었을 때 상용 모델 특유의 장황한 부연설명이나 배경지식의 혼입 없이 현장 친화적인 정보 밀도와 답변의 충실도를 유지했기 때문으로 분석된다. 결과적으로, 지식베이스가 체계적으로 관리되고 최적의 토픽이 LLM에 전달되는 구조 하에서는 오픈소스 모델이 상용 모델과 대등하거나, 도메인 특화 태스크에서 상대적으로 더 높은 정밀도를 구현할 수 있음을 시사한다.

Ⅲ. 결 론

본 연구는 기술-조직-환경(TOE) 프레임워크를 통해 제조 기업의 생성형 AI 도입 의사결정에 영향을 미치는 다차원적 요인을 규명하고, On-Premise RAG 시스템의 실효성을 보여줌으로써 다음과 같은 학술적 및 실천적 시사점을 제공한다.

첫째, 기술적 맥락에서의 이론적 시사점이다. 본 연구는 AI 모델의 효용성이 단순히 파라미터 규모나 범용 지식의 양에 비례한다는 기술적 통념을 실증 결과를 통해 재검토하고, 과업 적합성 중심의 새로운 해석을 제시하였다. 연구 결과, 상대적으로 적은 파라미터를 가진 오픈소스 모델이라도 고품질 검색 모듈과 결합된 도메인 특화 시스템으로 구성될 경우 거대 상용 모델과 대등하거나 정성적 측면(유용성, 간결성, 관련성)에서는 상대적으로 우수한 성과를 보였다. 이는 LLM의 추론 능력이 독립적으로 존재하는 것이 아니라 외부 지식 소스와의 유기적 상호작용 속에서 극대화됨을 의미하며, 향후 AI 연구가 개별 모델의 지능 자체보다는 목표 과업에 최적화된 시스템 아키텍처의 효율성을 규명하는 방향으로 확장되어야 함을 시사한다. 상용 LLM 기반 시스템이 갖는 환각 현상 대비 도메인 튜닝된 오픈소스 모델이 보여준 정보의 정확성과 간결성은 전문 도메인 환경에서 오픈소스 모델이 갖는 차별화된 상대적 이점에 대한 이론적 근거를 확립하였다.

둘째, 조직적 맥락에서의 실무적 시사점이다. 본 연구는 비용 효율성이라는 조직적 제약을 해소할 수 있는 현실적인 의사결정 기준을 마련하였다. 오픈소스 모델 RAG가 최소한의 토큰 비용으로도 상용 모델에 준하는 품질을 달성함을 보여주어, 지속적인 운영 비용을 절감할 수 있는 경제적 타당성을 확보하였다. 이는 기업이 내부 가용 인프라에 맞춰 모델을 유연하게 최적화할 수 있는 조직적 유연성을 제공하여 경영진의 도입 불확실성을 낮추는 데 기여한다.

셋째, 환경적 맥락에서의 정책적 시사점이다. 제조 데이터의 외부 유출 우려는 기업의 디지털 전환을 저해하는 가장 큰 환경적 장벽이다. 본 연구는 오픈소스 모델 기반의 On-Premise RAG가 이러한 환경적 제약을 기술적으로 극복할 수 있는 대안임을 실증하였으며, 이는 정부의 중소기업 지원 정책이 단순한 클라우드 바우처 제공을 넘어 설치형 경량화 모델의 보급을 지원하는 방향전환의 필요성이 제기됨. 나아가 기업 차원의 데이터 통제권 확보는 궁극적으로 특정 글로벌 빅테크 기업에 대한 국가적 기술 종속을 방지하고, 제조 강국 국가의 AI 기술 주권을 확보하는 토대가 될 것이다.

결론적으로 본 연구는 TOE 프레임워크의 다차원적 분석을 통해 On-Premise RAG가 보안이 중

요한 제조 전문 도메인에서 신뢰성과 비용 효율성을 동시에 확보할 수 있는 기술경영학적 대안임을 실증적으로 제시한다.

REFERENCES

[1] McKinsey & Company. (2025). *The state of AI in 2025: Agents, innovation, and transformation.*

[2] Tornatzky, L. G., & Fleischer, M., *The processes of technological innovation,* Lexington Books, 1990.

[3] Ji, Z., et al., *Survey of Hallucination in Natural Language Generation,* ACM Computing Surveys, Vol. 55, no. 12, pp. 1-38, 2023.

[4] Gao, Y., et al., *Retrieval-Augmented Generation for Large Language Models: A Survey*, arXiv preprint arXiv:2312.10997, 2023.

[5] Lewis, P., et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*,

[6] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). *Judging llm-as-a-judge with mt-bench and chatbot arena.* arXiv preprintarXiv:2306.05685.

[7] Abdulkhader, N. K., *Evaluating retrieval strategies for domain-specific RAG systems through automated metrics and human-centered user study*, [Master's thesis, University of Skövde, 2025].

[8] Lin, C.-Y., ROUGE: *A Package for Automatic Evaluation of Summaries, Text Summarization Branches Out*, pp. 74 - 81, Barcelona, Spain, 2004.

[9] Papineni, K., et al., BLEU: a method for automatic evaluation of machine translation, *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311 - 318, Philadelphia, USA, 2002

[10] Zhang, T., et al., BERTScore: Evaluating Text Generation with BERT, *In International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

저 자 소 개

유재석(정회원)

1999년 한림대학교 전자공학과 학사 졸업.

2025년 경희대학교 테크노경영대학원 AI기술경영학과 석사 대학원생.

2025년 텔스타 주식회사 연구소장

<주관심분야 : 스마트기술, 자동화 산업기술, AI 기술경영 >

홍아름(정회원)

2011년 서울대학교 기술경영경제정책협동과정 석박사 졸업.

2025년 경희대학교 테크노경영대학원 AI기술경영학과 부교수

<주관심분야 : 스마트기술, 산업기술정책, AI기술경영>